خوارزمية لضغط النصوص العربية

إعداد

د. حسن مبروك الصَّلّاي

د. يحي محمد الحاج

فهد حمود القحطاني

وحدة البحث العلمي بكلية علوم الحاسب والمعلومات

جامعة الإمام محمد بن سعود الإسلامية

ملخص

تعتبر تقنية ضغط البيانات من أهم تقنيات العصر، فقد أسهمت في تطور التقنية الحديثة تطوراً هائلا. حيث إنه يمكن تخزين عدد كبير من البيانات على قرص صغير جداً لا يتعدى حجمه بضعة سنتيمترات. هناك العديد من البرامج الحاسوبية تمكن من ضغط النصوص العربية إلا أنها لا تأخذ بعين الاعتبار خصائص اللغة العربية التي من أهمها كونها لغة اشتقاقية. تهدف هذه الورقة إلى تطوير خوارزمية ضغط خاصة باللغة العربية وتطوير بيئة حاسوبية مجانية ومفتوحة المصدر لضغط النصوص العربية أو للمعالجة الأولية لهذه النصوص للتقليل من حجمها لتضغط بعد ذلك بأحد برامج الضغط المعروفة عالميا.

الكلمات المفتاحية: ضغط البيانات، اللغة العربية، الضغط حسب موضوع النص، خوارزميات

مقدمة

لا يقلل من شأن اللغة العربية إلا جاهل بها. إنها ليست فقط اللغة الأم لأكثر من ١٥٠ مليون عربي مابين المغرب الأقصى والعراق بل تعتبر هذه اللغة لغة مقدسة في أعين المسلمين الذين يقدرون بخمس العالم وذلك لأن القرآن نزل بلسان عربي مبين. ولهذا فلا يمكن أبدا لهذه اللغة أن تتجاوز أو تهمل فهي محفوظة بحفظ الله عز وجل للقرآن الكريم. وهي من حيث الثراء اللغوي من أثرى لغات العالم وذلك لأنها من اللغات الاشتقاقية ومع هذا فإنها من اللغات الثابتة الناضجة التي لا تتغير مع الزمن إلا قليلا فالعربي اليوم يمكن له فهم القرآن أو الشعر الجاهلي وإن أصابته بعض العجمة بينما اللغات الأخرى مثل الإنجليزية أو الألمانية فيصعب على أهلها فهم لغة أسلافهم بحكم التطورات التي حصلت فيها فهي لغات ضعيفة الثبوت. من جهة أخرى فإن الثورة المعلوماتية التي عرفتها الإنسانية جعلت من الحاسب الآلي القلب النابض لمعظم الأعمال اليوم. فيندر أن تجد عملا من الأعمال مع اختلاف أنواعها وتباين أشكالها إلا وللحاسب الآلي فيه مكان خصوصا بعد الانفجار الهائل في كم تناقل المعلومات على مستو ى العالم مع ظهور الشبكات واسعة النطاق التي من أهمها أو أهمها على الإطلاق الإنترنت. هذا التطور السريع ولد الحاجة الماسة إلى جعل تبادل المعلومات وتخزينها على العتاد المادي أن يكون بشكل أسهل وأسرع وأقل كلفة. ومن هذا المنطلق فإن

تقنية ضغط البيانات تعتبر من أهم تقنيات العصر، فقد أسهمت في تطور التقنية الحديثة تطوراً هائلا. حيث إنه يمكن تخزين عدد كبير من البيانات على قرص صغير جداً لا يتعدى حجمه بضعة سنتيمترات. هناك العديد من البرامج الحاسوبية تقبل التعامل وضغط النصوص العربية إلا أنها لا تأخذ بعين الاعتبار خصائص اللغة العربية التي من أهمها أنها لغة اشتقاقية. تهدف هذه الورقة إلى تطوير خوارزمية ضغط خاصة باللغة العربية وتطوير بيئة حاسوبية مجانية ومفتوحة المصدر لضغط النصوص العربية أو للمعالجة الأولية لهذه النصوص للتقليل من حجمها لتضغط بعد ذلك بأحد برامج الضغط المعروفة عالميا. تبدأ هذه الورقة بإعطاء نبذة مختصرة عن تقنية ضغط البيانات وذلك بذكر أهم أنواعها وأهم خوارزميات الضغط المعروفة وأيضا أدوات وبرامج الضغط المشهورة ثم تطرح خوارزمية تأخذ بعين الاعتبار خصائص اللغة العربية مع دراسة أولية تقيس أداء هذه الخوارزمية.

تقنيات وطرق ضغط البيانات

تكمن أهمية تقنية ضغط البيانات في حالة تخزين المعلومات على العتاد المادي أو إرسالها عن طريق الشبكة. إذ أنه يمكن ضغط المعلومة بالحجم المناسب أو بمعدَل إرسال مناسب حسب طبيعة التطبيقات. ومن هنا يمكن التوفير في موارد تخزين البيانات وأيضا تخفيض ضغط استعمال موارد نقل البيانات على الشبكة مع تسريع عملية نقل هذه البيانات. عموما يمكن تقسيم طرق ضغط البيانات إلى فئتين: طرق الضغط دون فقدان أي جزء من البيانات المعدة للضغط Lossless Compression وطرق الضغط بفقدان بعض من البيانات Lossy Compression . تضمن طرق الضغط دون فقدان للمعلومات المضغوطة إمكانية إعادتها إلى صورها الأصلية بالضبط و يستخدم للبيانات المهمة كبعض أنواع الصور كالصور الطبية و الملفات التنفيذية exe, والملفات النصية txt, doc ....الخ.

ويتم استخدام طرق الضغط بفقدان بعض من البيانات عند الرغبة في الحصول على نسبة ضغط عالية جدا وليست هناك حاجه ضرورية لأن يكون الملف الناتج بعد عملية الضغط مطابقا تماما للملف الأصلي. إذ أنه في كثير من التطبيقات، النقص الحاصل في الإعادة الدقيقة للهيكلة غير مهم وبالإمكان تحمل (قبول) ضياع كمية كبيرة من المعلومات.

وهذا النوع من الضغط مناسب تماما للملفات الوسائط المتعددة (الملتيميديا) كملفات الصور وملفات الصوت وملفات الفيديو [4,3,1].

هناك عدد كبير من تقنيات الضغط أو خوارزميات ضغط الملفات تندرج تحت كلا من الأسلوبين السابق ذكرهما مثل Run Length Encoding, Huffman Coding [2], Arithmetic Coding, LZ-77 Encoding [6,7], LZH, LZW Coding [5,12] بالنسبة للضغط دون فقدان و GIF و JPEG، وWavelet Fractal، وMP3 و MPEG-4،Asf للصور الثابتة والصوت والفيديو بالنسبة للضغط بفقدان بعض البيانات.

ومن أشهر البرامج المستخدمة حالياً للضغط دون فقدان بعض البيانات برامج Winzip [10] و WinRAR[11] وبرنامج JBEG المستخدم للفاكس التي تعمل على نظام windows وهي غير مفتوحة المصدر و برامج Compress و Gzip و 7zip [13] المجانية ومفتوحة المصدر والتي تعمل على نظام Unix. ويعتبر برنامج Winzip الأشهر والأسرع أيضا في عملية الضغط إلا أن برنامج WinRAR يتميز بمعدلات ضغط أقوى ولكنه أقل سرعة من Winzip. تدمج هذه البرامج خوارزمية Huffman مع خوارزمية LZW.

كل هذه البرامج وإن كانت تقبل التعامل وضغط النصوص العربية إلا أنها لا تأخذ بعين الاعتبار خصائص اللغة العربية التي من أهمها أنها لغة اشتقاقية.

خوارزمية لضغط النصوص العربية

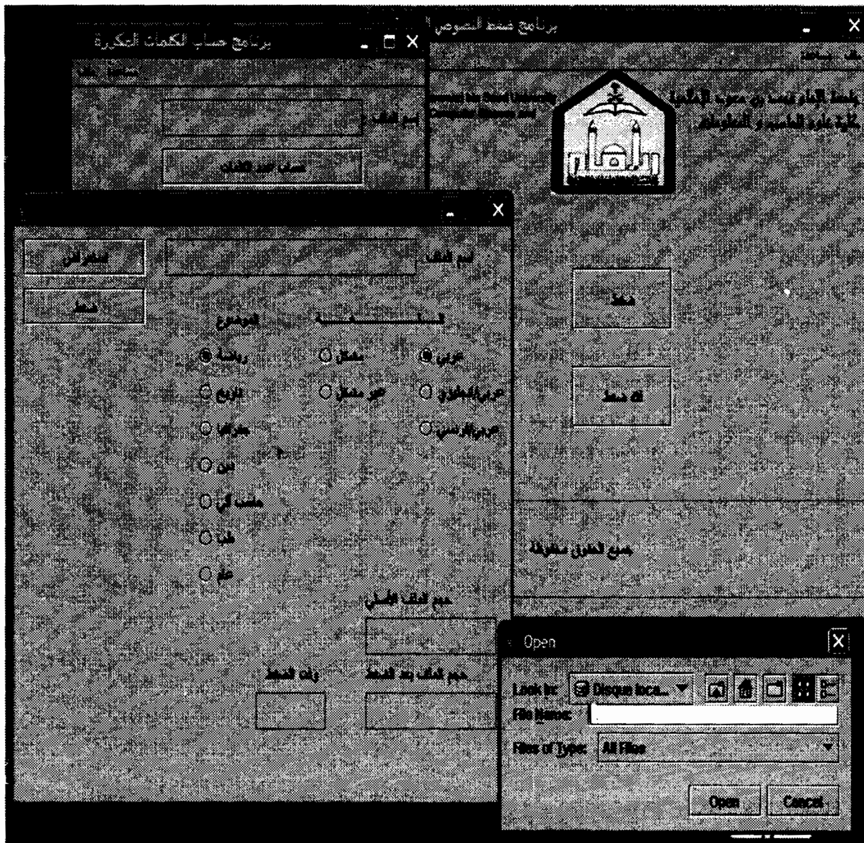
نتناول في هذه الفقرة بشيء من الإجمال الخوارزمية المقترحة التي تعتمد على بعض خصائص اللغة العربية الأساسية وللتفصيل الدقيق يمكن الرجوع إلى [٩].

تعتمد الخوارزمية على فكرة الضغط بتعويض أكثر الكلمات استعمالا في اللغة العربية برموز الحروف الأجنبية و بعض رموز الآسكي غير المستعملة في النصوص العربية. ثم تمرير الناتج بعد هذه المعالجة الأولية للنص إلى إحدى أدوات أو برامج الضغط المعروفة لنحصل على نسبة ضغط أعلى من لو أننا ضغطنا نفس النص بنفس البرنامج من غير هذه المعالجة الأولية. في عملية فك الضغط نقوم باستعمال نفس برنامج الضغط للحصول على النص المختلط (عربي+رموز الكلمات) ومن ثم استرجاع النص الأساسي بتمرير معالج الرموز لاسترجاع النص الأصلي. تشبه هذه الخوارزمية خوارزمية LZW إلا أنها تقوم بترميز الكلمات على ٨ بت لا على ١٢ بت كما في LZW وهي أفضل من هذه الناحية إلا أن محدودية الكلمات والحروف الأجنبية المرمزة على ٨ بت تقف عائقا أمام هذه الخوارزمية لتصل إلى قدرة عمومية التعامل مع جميع النصوص كما هو الحال في خوارزمية LZW. تتجاوز الخوارزمية هذا المشكل بجعل الرمز يتكون من أكثر من حرف مع ضمان التغاير بين الرموز. كلما تعتمد الخوارزمية أيضا على فكرة الضغط حسب الموضوع

أي قبل عملية الضغط يحدد المستخدم طبيعة النص من حيث الشكل (مشكل، غير مشكل، يحتوي على حروف أجنبية أم لا ...) والمحتوى (دين، تاريخ، جغرافيا، هندسة، اقتصاد، رياضة، عام...) ويتم تخصيص قوائم للكلمات الأكثر استعمالا في كل محتوى وقوائم خاصة بالسوابق واللواحق في اللغة العربية كيما يكون الضغط أكثر فاعلية. لو قدر وجود نفس الرموز الأجنبية في نفس النص المضغوط فإن الخوارزمية تمر في المرة الأولى على النص بمجمله لتحصر هذه الرموز وتستعمل بقية الرموز في الضغط كما تضيف رمزا خاصا قبل وبعد كل سلسلة من الرموز الأجنبية لكي تتعرف عليها عند عملية فك الضغط.

البرمجة والقياس الأولي لأداء الخوارزمية

تم برمجة الخوارزمية مع واجهة برنامج الضغط الخاص باللغة العربية (انظر شكل رقم١) باستعمال لغة JAVA لنضمن اشتغال البرنامج على أكثر من نظام تشغيل ولكن من أهم سلبيات JAVA كما هو معروف استهلاك موارد الذاكرة والبطء في التنفيذ ونأمل في مرحلة مقبلة برمجة أداة الضغط بلغة Assembly لتسريع عملية الضغط.

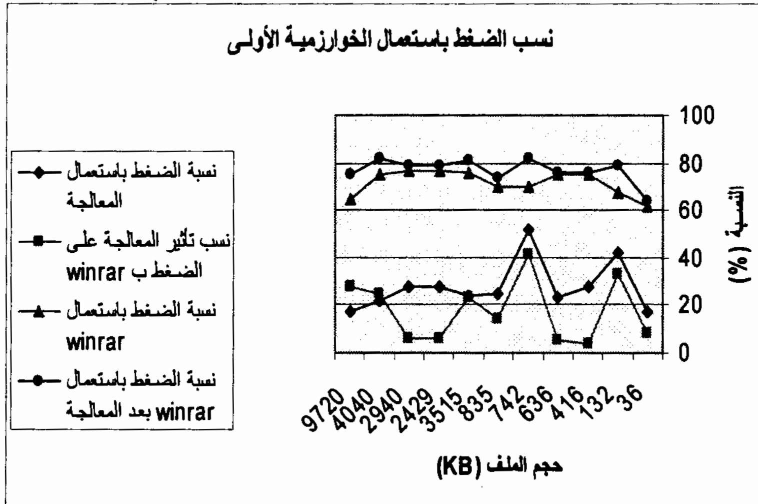


الشكل ١: واجهة النظام التي تحتوي على الخيارات الأساسية.

قمنا بحساب الكلمات الأكثر استعمالا في بعض كتب الحديث وإدراج بعض السوابق واللواحق للغة العربية بمجموع ٢٠٠ عنصر ثم قمنا بقياس نسبة الضغط باستعمال الخوارزمية وتمرير برنامج winrar للنصوص قبل معالجتها بالخوارزمية وبعد المعالجة مع قياس المدة الزمنية التي تستغرقها المعالجة الأولية على الكتب التالية (انظر جدول رقم 1)

الملف	الحجم الأصلي (KB)	الحجم بعد المعالجة (KB)	الحجم ب winrar بدون المعالجة (KB)	الحجم ب winrar بعد المعالجة (KB)
شرح السنة للبربهاري	36	30	14	13
الإبانة لابن بطة	132	77	43	29
مسند الشافعي	416	301	104	100
الشريعة للآجري	636	490	165	157
الأذْكَارُ النَّوَوِيَّة للنَّوَوي	742	359	226	134
شرح كتاب الإيمان من مسلم	835	631	257	222
معرفة السنن والآثار للبيهقي	3515	2706	868	672
صحيح مسلم	2429	1750	562	533
صحيح البخاري	2940	2130	684	645
المنهاج شرح مسلم	4040	3170	1010	767
سير أعلام النبلاء	9720	8120	3455	2520

جدول رقم 1: مجموعة الاختبار وقياس أداء الخوارزمية



الشكل ٢: نسب الضغط باستعمال الخوارزمية.

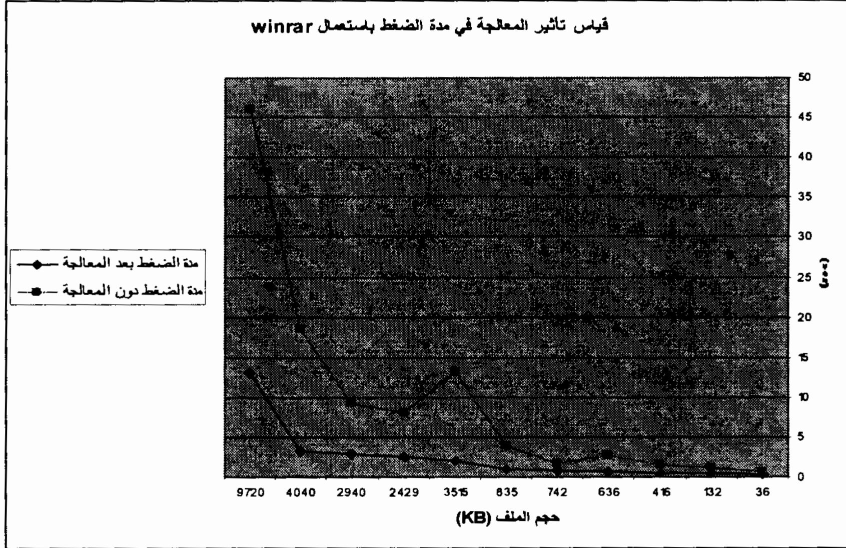
يظهر الشكل رقم٢ رسما بيانيا يعكس نسب الضغط باستعمال الخوارزمية التي يلخصها الجدول رقم ٢. فنلاحظ أن الخوارزمية، وبالنسبة إلى مجموعة الاختبار التي استعملناها والتي تركز على كتب الحديث بصفة خاصة، توفر نسب ضغط تصل في المتوسط العام إلى ٢٨% وفي أقصاها إلى ٥٢%

	بعد المعالجة الأولية	winrar	winrar + المعالجة الأولية	تأثير الخوارزمية على الضغط بwinrar
متوسط نسب الضغط %	28	72	77	18
أدنى نسب الضغط %	17	62	64	4
أعلى نسب الضغط %	52	77	82	41

جدول رقم ٢: قياس نسب الضغط

وترفع أداء winrar بمعدل ١٨% في المتوسط وتصل إلى ٤١% في أفضل الحالات كما ترفع نسبة الضغط لwinrar ب٥% تقريبا كما

يمكن الرفع من مستوى نسب الضغط بزيادة عدد الكلمات والسوابق واللواحق الأكثر استعمالا في اللغة العربية.



الشكل ٣: تأثير المعالجة بالخوارزمية على مدة الضغط بwinrar.

إلا أن هذه المعالجة الأولية تزيد في مدة الضغط بحسب حجم الملف المضغوط وطبيعة الكلمات فيه (انظر الشكل رقم ٣) فملف بحجم ١٠MB يكلف ٤٦ ثانية في مقابل ١٣ ثانية بدون المعالجة الأولية ويرجع هذا بالأساس إلى استعمالنا للغة JAVA في البرمجة ويمكن تحسين الأداء باستعمال لغة C أو Assembly.

الخاتمة

تقدم هذه الورقة خوارزمية لضغط النصوص التي ترتكز على بعض خصائص اللغة العربية. يتيح البرنامج حساب الكلمات الأكثر استعمالا حسب المواضيع كما يتيح الضغط وفك الضغط باستعمال الخوارزمية. لا يتعامل البرنامج في المرحلة الحالية إلا مع ملفات txt. ونأمل في المستقبل استكمال بقية الجوانب البرمجية وتطوير البرنامج للتعامل مع بقية أنواع الملفات المشهورة. الدراسة الأولية لأداء الخوارزمية تعتبر مشجعة ويدعونا للمزيد من العمل الذي لا يزال في بدايته ويحتاج إلى جهد كبير خصوصا على مستوى الجانب البرمجي والله الموفق لكل خير.

المراجع

1. Abramson, N. 1963. Information Theory and Coding. McGraw-Hill, New York.
2. Cormack, G. V., and Horspool, R. N. 1984. Algorithms for Adaptive Huffman Codes. Inform. Process. Lett. 18, 3 (Mar.), 159-165.
3. Fraenkel, A. S., Mor, M., and Perl, Y. 1983. Is Text Compression by Prefixes and Suffixes Practical? Acta Inf. 20, 4 (Dec.), 371-375. Wilkins, L. C., and Wintz, P. A. 1971.
4. Bibliography on Data Compression, Picture Properties and Picture Coding. IEEE Trans. Inform. Theory 17, 2, 180-197.
5. Welch, T. A. 1984. A Technique for High-Performance Data Compression. Computer 17, 6 (June), 8-19
6. Ziv, J., and Lempel, A. 1977. A Universal Algorithm for Sequential Data Compression. IEEE Trans. Inform. Theory 23, 3 (May), 337-343.

7. Ziv, J., and Lempel, A. 1978. Compression of Individual Sequences via Variable-Rate Coding. IEEE Trans. Inform. Theory 24, 5 (Sept.), 530-536.
8. Mohamed Y. Osman and Mohammed Al-Habib, Arabic Text Compression Using Huffman Code, The Arabian Journal for Science and Engineering, Volume 16, Number 4B.
9. Sallay, H. and Qahtani F. Y 2007, An Arabic Compression Tool, Technical report.
10. http://www.winzip.com/
11. www.win-rar.com
12. http:// www. dogma. net/markn/articles/lzw/lzw .htm
13. http://en.wikipedia.org/wiki/7-Zip.